.

Overview

The following is an excerpt from Sigman, Richard S. (2000), "Estimation and Variance Estimation in a Standardized Economic Processing System, " *Proceedings of the International Conference on Establishment Surveys*, Alexandrai, VA: American Statistical Association, pp. 677-686:

Like other StEPS modules, the following are the major components of the StEPS Estimates and Variances Module:

- Standard data set structures for micro data, macro data, and processing parameters;
- Interactive screens for specifying parameters, submitting batch jobs, and requesting results listings; and
- SAS macros and scripts for batch calculations.

Each of these is discussed below.

Standard data set structures

StEPS stores micro data in *control files* and *item files*. Micro data includes data associated with questionnaire items; data associated with survey operations such as sample selection, mailing, collection, or check-in; or auxiliary data available from censuses or administrative sources. The item file can contain only numeric micro data, whereas the control file can contain numeric and character data. Another difference between the control file and the item file is that the control file has a "fat" format, whereas the item file has "skinny" format. In the control file (i.e., fat format) there is one record per reporting unit (ID), and the fields within each record correspond to control-file variables. In the item file (i.e., skinny format) there is one record per ID/item combination, and fields within each record correspond to different *data versions* (plus there is a field containing a data flag).

StEPS stores the following data versions in each record of the item file:

r_{ij} = reported data for item i and reporting unit j
e_{ij} = edited data for item it and reporting unit j
a_{ij} = adjusted data for item it and reporting unit j
w_{ij} = weighted-adjusted data for item it and reporting unit j

The default value for edited data is $e_{ij} = r_{ij}$. StEPS users, however, may change edited data by using the Review and Correction Module, or StEPS can change edited data via the Imputation Module.

Some surveys adjust micro data for data collection effects, such as trading day effects in monthly surveys or in annual surveys the effect on reported inventories of ending inventory dates other than December 31. One way that StEPS adjusts micro data is

$$a_{ij} = f(t_i, \mathbf{B}_j)\, e_{ij} ,$$

where

t_i = the value for item i of a variable, called *adjustment type* stored on the *item data dictionary* file;
$\mathbf{B}_j$ = a vector of *BY variables* --i.e., categorical variables--associated with reporting unit j; and
f( ) = a SAS format that StEPS creates to map the vector $(t_i, \mathbf{B}_j)$ into user-provide adjustment factors.

Another way StEPS adjusts micro data is to use user-provided SAS code stored in the *adjust/derive definitions file*. Many surveys do not adjust their micro data, however, in which case $a_{ij} = e_{ij}$.

StEPS calculates weighted-adjusted data using the following formula:

$$w_{ij} = \omega_j\, g_{n(i),j}\, a_{ij} .$$

The quantity ω_j is the sampling weight for reporting unit j. The control file stores three *g weights*, g_{1i}, g_{2i}, and g_{3i}, for each reporting unit. We had planned to use the g-weights in the manner described in Estavao, et al. (1995), in which if they are chosen appropriately the resulting weighted totals (or weighted means) are generalized regression estimators. To date, we have not used the g-weights for this purpose. One way we have used the g-weights was in our annual retail trade survey, which collects some items for only a subsample of the survey, we let ω_j store the first-phase sampling

weight and let the g-weight store the second-phase weight. In the future we plan to use the g-weights to store non-response adjustment factors for surveys that use weight adjustment to handle unit nonresponse. The g-weight is equal to 1.0 for unweighted and Horvitz-Thompson estimators. The quantity *n(i)* is the *g-weight number* and indicates which g-weight, g_{1i} , g_{2i} , or g_{3i}, is associated with item i. If *n(i)*=0 then item i has a g-weight of 1.0. The g-weight number, like the adjustment type, t_i, is stored in the item data dictionary, which contains one record for each item-data variable.

The item file's skinny format can be difficult to use for estimation and variance calculations. Consequently, StEPS can create an *estimation fat file*, which has one record per ID, and the fields within each record can be any of the following: control file variable; adjusted or weighted-adjusted version of an item file variable; constant data; or *recode*, which is a variable created at the time of fat-file creation via a user-provided SAS expression involving other fat-file variables. When StEPS creates an estimation fat file, a variable on the control file, called the *weighting switch,* selects for each ID the adjusted or weighted-adjusted version of the item file variables. Certain values of the weighting switch zero out item data in the fat file or delete an entire record from the estimation fat file. By setting the weighting switch to a particular value for each ID, one can control the contents of each estimation-fat-file record, for purposes such as handling deaths by zeroing out or deleting data or handling outliers by deleting or down-weighting to self-representing.

StEPS stores macro data in *estimation results files* (ERFs). One ERF corresponds to one *table*, which is the result of StEPS performing calculations on *analysis variables* for individual values of categorical *BY variables.* The types of results StEPS stores in ERFs include: totals, ratios, trends, other derived estimates (i.e., functions of totals), standard errors, CVs, covariances, t-tests, imputation rates and disclosure-avoidance information. The ERF has a skinny format--each ERF record contains only one calculated result, with other variables in the record identifying the type of result, the name(s) of the analysis variable(s), and the value(s) of any BY variable(s).

Two files store estimation processing information: the *estimation specification file* (ESF) and the *estimation formulas file* (EFF). The ESF stores parameters used by the SAS macros described in section 5.3; the EFF stores SAS expressions and SAS code, also used by the SAS estimation macros. Both the ESF and EFF are populated via interactive screens. Developing a file layout for the ESF was challenging. We rejected a skinny-record format of one record per parameter because of the complexity of file updating from screens displaying multiple parameters. Instead, we decided the ESF would have one record per *specification*, which is a vector of parameters displayed together on the same screen. In the ESF, sets of specifications (i.e., records) associated with the same type of screen and processing action are called *objects*. For example, the "BY object" contains specifications for BY variables, whereas the derived object contains specifications for the calculation of derived estimates.

Interactive screens

Interactive screen in the Estimates and Variances Module allow StEPS users to do the following:

- Calculate weighted data for all items and IDs in the item file.
- Run Quicktab program, which calculates weighted totals, year-to-year trends, imputation rates, unweighted counts, and disclosure-avoidance information. The Quicktab program requires analysis variables to be item file variables and any BY variables to be control file variables. Quicktab does not calculate standard errors, CVs, or derived estimates. The possible outputs from Quicktab are a SAS data set, an ASCII file (for downloading), printer output, or the SAS Output Window.
- Enter and modify specifications and formulas for use by batch jobs. Specifications and formulas tell StEPS WHAT to estimate with WHAT data. The StEPS user can select analysis and BY variables (from item data, control data, recodes, or constants); specify the method of calculating standard errors (random groups, VPLX replication, or formulas for Poisson or Tillé samplig); enter formulas for derived estimates and the derivatives of non-linear estimators); copy results from one ERF to another ERF; and remove results from an ERF.
- Submit estimation scripts to run in batch. Scripts are described in more detail in section 5.3. A screen displays the available scripts, and the user selects one of the displayed scripts to run immediately or at a scheduled time.
- Review estimation results. A screen displays a list of ERFs, and the user can select an ERF for interactive

viewing with SAS/FSVIEW® or for formatting by StEPS into a printed listing.

SAS macros and scripts

StEPS scripts execute SAS code that is part of StEPS or has been generated by StEPS. For the Estimates and Variances Module, scripts execute SAS code that is part of StEPS. In particular, estimation scripts execute one or more of the following SAS macros:

%extract — Creates estimate fat file.
%totals — Calculates totals and imputation rates.
%derive — Calculates derived estimates.
%erfrmt — Reformats an ERF.
%rtsumvar — Aggregates totals, standard errors, and imputation rates.
%copy1 — Copies results between ERFs.
%remove — Removes results from an ERF.
%round — Rounds totals and standard errors
%vpl2stp — Stores VPLX-calculated estimates and standard errors in an ERF (Dajani 1999a).\
%rgvar_1 — Calculates replicate totals from random group totals.
%rgvar_2 — Calculates replicate-based standard errors from replicate estimates.
%vrncs_p — Calculates standard errors for Poisson-sampling designs.
%vrncs_t — Calculates standard errors for Tillé sampling designs.
%cvrncs_t — Calculates covariances for Tillé sampling designs.
%taylor — Calculates standard errors of non-linar estimates using Taylor approximation.

Many of these macros are individually controlled by parameters analysts have entered into the ESF and EFF. Parameters specify WHAT to estimate and WHAT data to use. The estimation script controls the overall logic of HOW to calculate estimates and variances. Because this depends on the sample design, surveys with different sample designs require different scripts. Also, the sample designer should be involved in developing an estimation script--either as an advisor or as the person who produces the script.